Statistical Methods for Data Science

Mini Project #2

Yagna Srinivasa Harsha Annadata

Yxa210024

Problem 1:

A. Table 1 represents the frequencies and proportions of categories in Maine and the figure 1 is the bar graph representation of Maine.
From the graph we can infer that there are more runners from Maine than from anywhere.
Figure1:

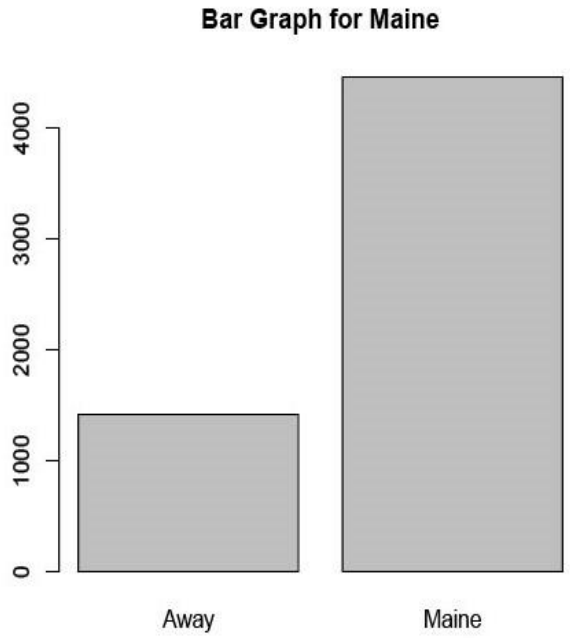


Table 1:

	Maine	Away
Count	1417	4458
Proportion of Table	0.2411	0.7588

B. Figure 2 represents the histograms for runner's time from Maine and Figure 3 represents the histogram for runner's time who are not from Maine. Table 2 summarizes all the runner's time.

We can observe that the distribution is symmetric.

Figure2:

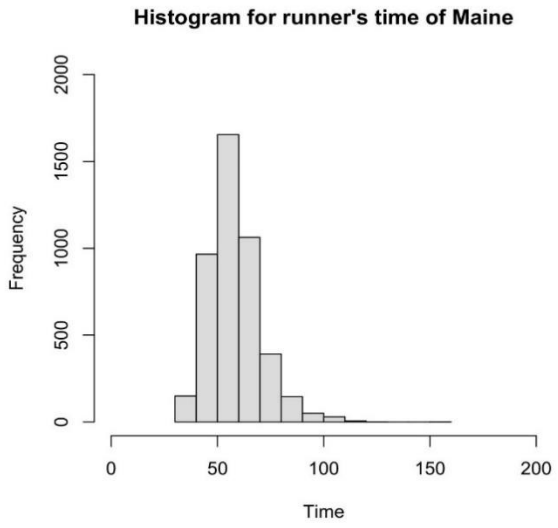


Figure3:

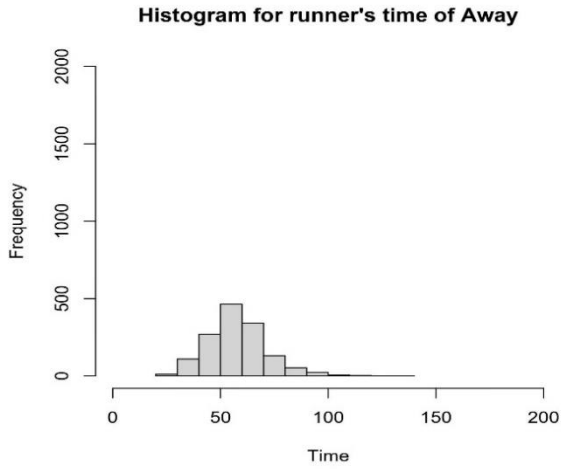
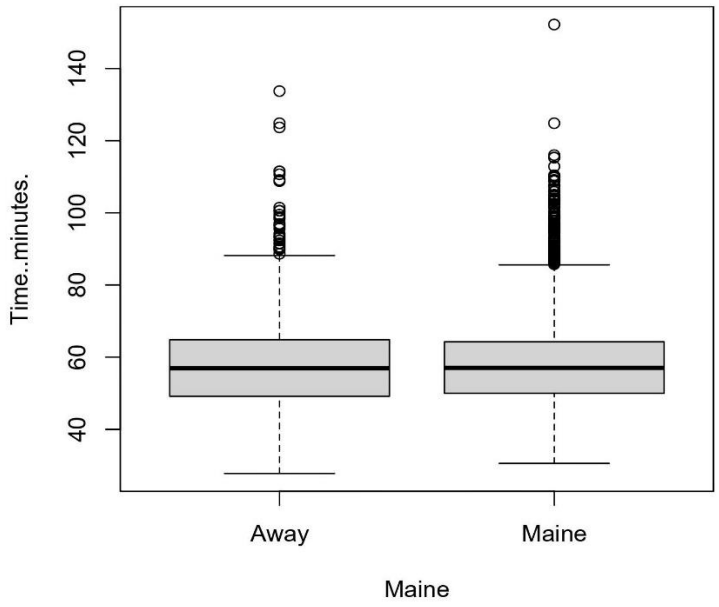


Table2:

	Min	Q1	Median	Mean	Q3	Max	IQR
Maine	30.57	50.00	57.03	58.20	64.24	152.17	14.247
Away	27.78	49.15	56.92	57.82	64.83	133.71	15.674

C. Figure 4 is the box plot for the runner's time from Maine and who are not from Maine. Q1, median and Q3 are similar. Distribution is symmetric.
Figure4:



D. Figure 5 is the boxplot for summary of Male and Female. Table 3 is the summary of the runner's age by gender. Q1, median and Q3 are larger for male than the female. Male boxplot is left skewed whereas female boxplot is right skewed.

Figure 5:

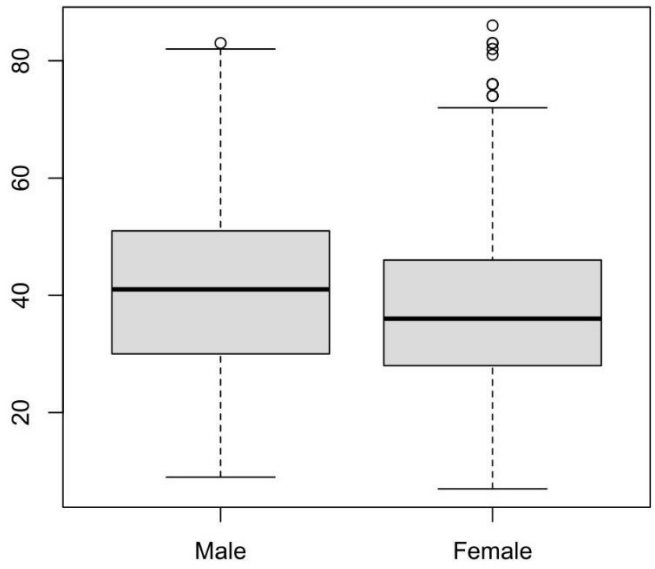


Table3:

	Min	Q1	Median	Mean	Q3	Max	IQR
Male	9.00	30.00	41.00	40.45	51.00	83.00	21.00
Female	7.00	28.00	36.00	37.24	46.00	86.00	18.00

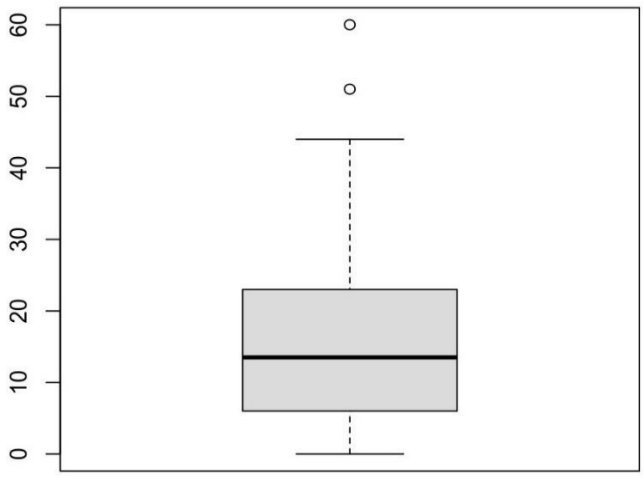
Problem 2:

Table 4 represents the summary for motorcycle accidents and Figure 6 is the box plot of motorcycle accidents. Motorcycle accident distribution is right skewed. There are 0 accidents in some states. Greenville and Horry have unusual motorcycle accidents. They have the highest. ¾ of the accidents are above 6 and ¼ are above 23 as observed from the table.

Table4:

Min	Q1	Median	Mean	Q3	Max	IQR
0.00	6.00	13.50	17.02	23.00	60.00	17.00

Figure6:



Section 2 R Code:

Problem 1:

```
R Console Page 1
```

Reading the data given in roadrace.csv

```
> RoadRace <-
read.csv("C:\\Users\\yxa210024\\Desktop\\Masters\\spring2023\\Stats for
DS\\mini_project2\\roadrace.csv", na.strings = "*")
> attach(RoadRace)
The following objects are masked from RoadRace (pos = 3):
Age, Division, Division.Entrants, Division.Place, From.USA, Maine,
Mile.pace..seconds., Place, Sex, State.Country, Time..minutes.,
Time..seconds.
The following objects are masked from RoadRace (pos = 4):
Age, Division, Division.Entrants, Division.Place, From.USA, Maine,
Mile.pace..seconds., Place, Sex, State.Country, Time..minutes.,
Time..seconds.
The following objects are masked from roadrace (pos = 6):
Age, Division, Division.Entrants, Division.Place, From.USA, Maine,
Mile.pace..seconds., Place, Sex, State.Country, Time..minutes.,
Time..seconds.
```

```
The following objects are masked from roadrace (pos = 7):
Age, Division, Division.Entrants, Division.Place, From.USA, Maine,
Mile.pace..seconds., Place, Sex, State.Country, Time..minutes.,
Time..seconds.
> colnames(RoadRace)
[1] "Place" "Division.Place" "Division.Entrants"
[4] "Division" "Age" "Sex"
[7] "State.Country" "Time..seconds." "Mile.pace..seconds."
[10] "From.USA" "Maine" "Time..minutes."
```

Plotting Bar graph and summary

```
> barplot(table(Maine), main = "Bar Graph for Maine")
> table(Maine)
Maine
Away Maine
1417 4458
> prop.table(table(Maine))
Maine
Away Maine
0.2411915 0.7588085
> M <- subset(RoadRace , Maine == "Maine")$Time..minutes.
> A <- subset(RoadRace , Maine == "Away")$Time..minutes.
> summary(M)
Min. 1st Qu. Median Mean 3rd Qu. Max.
30.57 50.00 57.03 58.20 64.24 152.17
> summary(A)
Min. 1st Qu. Median Mean 3rd Qu. Max.
27.78 49.15 56.92 57.82 64.83 133.71
> IQR(M)
[1] 14.24775
> IQR(A)
[1] 15.674
```

Plotting Histogram for Maine runner's time

```
> hist(M, xlim = c(0, 200), ylim = c(0,2000), xlab = "Time", main =
"Histogram for runner's time
of Maine")
> hist(A, xlim = c(0, 200), ylim = c(0,2000), xlab = "Time", main =
"Histogram for runner's time
of Away")
```

Plotting box graph for the Maine runner's time

```
> boxplot(Time..minutes.~Maine)
```

Plotting box graph for male and female runners with summary

```
> Male <- Age[Sex == "M"]
> Female <- Age[Sex == "F"]
> boxplot(Male, Female, names = c("Male", "Female"))
> summary(Male)
Min. 1st Qu. Median Mean 3rd Qu. Max.
9.00 30.00 41.00 40.45 51.00 83.00
> summary(Female)
Min. 1st Qu. Median Mean 3rd Qu. Max.
7.00 28.00 36.00 37.24 46.00 86.00

>IQR(Male)
[1] 21
> IQR(Female)
[1] 18
```

Problem 2:

```
R Console Page 1
```

Reading the data given in motorcycle.csv

```
> MotorCycleAccidents <-
read.csv("C:\\Users\\yxa210024\\Desktop\\Masters\\spring2023\\Stats for
DS\\mini_project2\\motorcycle.csv")
> attach(MotorCycleAccidents)
The following objects are masked from motor:
County, Fatal.Motorcycle.Accidents
```

Plotting box graph for fatal motorcycle accidents

```
> boxplot(Fatal.Motorcycle.Accidents)
```

Finding the outliers

```
> BoxGraph <-boxplot(Fatal.Motorcycle.Accidents)
> BoxGraph$out
[1] 51 60
```

Displaying all the data

```
> tail(MotorCycleAccidents[order(Fatal.Motorcycle.Accidents), ], 100)
County Fatal.Motorcycle.Accidents
47 OTHER 0
48 UNKNOWN 0
1 ABBEVILLE 3
3 ALLENDALE 3
5 BAMBERG 3
19 EDGEFIELD 3
33 MCCORMICK 3
41 SALUDA 3
25 HAMPTON 5
36 NEWBERRY 5
44 UNION 5
9 CALHOUN 6
17 DILLON 6
6 BARNWELL 7
20 FAIRFIELD 7
35 MARLBORO 8
45 WILLIAMSBURG 10
11 CHEROKEE 11
37 OCONEE 11
13 CHESTERFIELD 12
24 GREENWOOD 12
27 JASPER 12
34 MARION 12
7 BEAUFORT 13
12 CHESTER 14
31 LEE 14
15 COLLETON 17
16 DARLINGTON 17
22 GEORGETOWN 17
29 LANCASTER 17
14 CLARENDON 18
28 KERSHAW 18
18 DORCHESTER 20
39 PICKENS 20
30 LAURENS 21
43 SUMTER 23
46 YORK 23
```

```
2 AIKEN 28
21 FLORENCE 29
38 ORANGEBURG 29
42 SPARTANBURG 30
32 LEXINGTON 34
4 ANDERSON 35
8 BERKELEY 38
40 RICHLAND 40
10 CHARLESTON 44
23 GREENVILLE 51
26 HORRY 60
```

{The highlighted red color details are the outliers}

Summary for Fatal motorcycle accidents

```
> summary(Fatal.Motorcycle.Accidents)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 6.00 13.50 17.02 23.00 60.00

> IQR(Fatal.Motorcycle.Accidents)
[1] 17
```